Sture Allén, som var forhindret i at deltage i mødet, har sendt denne artikel, som stort set dækker, hvad han ville have sagt i sit foredrag.

STURE ALLÉN, som är född i Göteborg 1928,
avlade fil kand-examen 1954, fil mag-examen 1956 och fil lic-examen 1961 samt disputerade för fil doktorsgrad på avhandlingen "Grafematisk analys som grundval för textedering med särskild hänsyn till Johan Ekeblads brev till brodern Claes Ekeblad 1639–1655" (Nordistica Gothoburgensia 1–2) 1965, allt i Göteborg. Förordnad till docent i nordiska språk vid Göteborgs universitet 1965. Tjänstledig för att leda projektet "Datamaskinell undersökning av tidningsprosa" med anslag från Riksbankens jubileumsfond från 1966. Förordnad till t f professor i nordiska språk vid Göteborgs universitet periodvis 1967–69. Förordnad till innehavare av en forskartjänst i språklig datateknik vid Statens humanistiska forskningsråd 1970. Utnämnd till professor i språklig databehandling vid Statens humanistiska forskningsråd 1972. Tilldelades Henrik Ahrenbergs pris av Göteborgs universitet 1966 och erhöll Svenska Akademiens språkvårdspris 1979. Invaldes i Kungliga Vetenskaps- och Vitterhets-Samhället i Göteborg 1976.

Utnämnd att fr o m den 1 juli 1979 vara professor i språkvetenskaplig databehandling vid universitetet i Göteborg.

Språkvetenskaplig databehandling

Frågorna från olika håll blir många, efterhand som ett nytt ämne växer fram. De kan ha sin utgångspunkt i att det hela verkar obegripligt eller fascinerande eller överflödigt eller nyttigt eller något annat. I allmänhet bottnar de i ett intresse att få veta mer. Undringarna tar här gestalt i en *kursiv* samtalspartner. Det utspinner sig ett belärande samtal eller som det en gång hette en dialogus.

Hur började det?

I vårt fall har ämnet spirat ur forskning i nordiska språk med intresse för frågor av allmänt språkvetenskaplig natur. Konkret gällde det i början att utveckla metodik för att snabbt få fram information om vilket element som helst i en text, t ex beläggen på ett ord eller en bokstav. Behovet kunde tillgodoses av en datamaskinell konkordans. Konkordansen till Johan Ekeblads brev togs fram 1961–62.

Sedan dess har en rad mer raffinerade konkordanstekniker skapats. Konkordanser av växlande slag är fortfarande det mest efterfrågade forskningshjälpmed-

Språkvetenskaplig databehandling
Sture Allén
Proceedings of NODALIDA 1979, pages 171-179

let både utom och inom institutionen. Här är några rader ur en konkordans över ordledet *log* med de fulla orden som kontext. Markeringarna till höger har att göra med klassifikation och beläggställe.

bi o	log	isk	AV -T	7306
bok kata	log		NN -EN	7944
chef s id eo	log		NN -EN	10224
Chicago soci o	log	i	NN -N	10262
dia	log		NN -EN	11807
dia	log	is er a nde	AV -=	11811
dia	log	scen	NN -EN	11815
djup psyk o	log	i	NN -N	12211
djur fys io	log	i	NN -N	12231
eko	log		NN -EN	13592
ent myt o	log	is er a	VB -AD	14182
epi	log		NN -EN	14245

Det finns också andra utgångspunkter för ämnet. På 1950-talet hade datamaskinen bland annat använts på olika håll för försök med språköversättning och informationssökning och för beräkningar i samband med textattribution eller författarbestämning.

Frekvensundersökningar nämns ofta. Vilken roll spelar de?

De är inte så dominerande som många tror. Jag menar då frekvensundersökningar som går utöver den rena bokföringen av teckensträngar mellan blanka typrum i en text, i den här meningen alltså "Jag", "menar", "då" osv. Mer avancerade frekvensundersökningar kräver utveckling av välspecificerade beskrivningsmodeller, något som leder rakt in i centrala lingvistiska problem. Märk väl att frekvensstudier inte alls nödvändigtvis gäller ord – som redan det är ett svårfångat begrepp – utan också ordförbindelser, ordled, grundbetydelser eller kärnbetydelser, syntaktiska konstruktioner och åtskilligt annat.

Det är fullt klart att en frekvensundersökning av modernt svenskt material, bedriven på det antydda sättet, har varit av grundläggande betydelse för ämnets framväxt här. De teoretiska och metodologiska övervägandena och tillämpningen av de utarbetade modellerna har kommit konturerna till ämnesområdet att framträda. Naturligtvis har internationella kontakter genom konferenser, arrangemang av forskarkurser och arbetsseminarier, samarbetsprojekt, publikationer och annat också givit betydelsefulla impulser.

Det projekt som har resulterat i Nusvensk frekvensordbok baserad på tidningstext skisserades 1964. Materialet samlades in 1965. Det bestod av färdigstansade hålremsor från tidningssätterier. Bearbetningarna började 1966 och utfördes inom den då bildade forskningsgruppen för modern svenska. Det är alltså i hög grad fråga om ett lagarbete. Den fjärde och sista ordboksdelen, som just gäller ordled och kärnbetydelser, produktionskörs hösten 1979. Formellt upplöstes forskningsgruppen i samband med att institutionen bildades 1977.

Hur kan man då karakterisera ämnet?

Språkvetenskaplig databehandling är som namnet anger ett språkvetenskapligt ämne. Det allmänna målet för ett sådant är att vinna insikt i naturligt språk och det vetenskapliga studiet av naturligt språk. Inom vårt speciella område är det grundläggande att betrakta språket som en process. Allmänt sett är processer mer komplicerade än strukturer genom att man måste ta hänsyn till ordningsföljd. Exempel på frågor som aktualiseras är hur man skall finna en satsdels gränser, hur man skall komma åt en diskontinuerlig förbindelses delar – ett exempel är *på* ett på förhand givet *sätt;* hur man skall identifiera och sammanföra böjningsformer och variantformer, hur man skall känna igen en lexikalisk huvudbetydelse – ***kursiv*** adjektiv 1 lutande (om stil) 2 mindre noggrann, oförberedd (om läsning och översättning) 3 fortlöpande, obegränsat pågående (om verbhandling); i vilken ordning man skall generera och presentera de olika konstituenterna i en struktur osv.

Denna dynamiska inriktning kräver lexikaliska och grammatiska komponenter som är långt mer detaljerade, täckande och välstrukturerade än vad som nu står till buds. De måste med ett ord vara explicita. Processynen leder enligt min tanke till uppställande av ett långsiktigt mål för forskningen inom ämnet som kan sammanfattas i formuleringen datorn som språkbrukare.

Datorn som språkbrukare?

Låt oss försöka undvika missförstånd. Den oförlikneliga språkbrukaren är människan. Det är vi överens om. Analogin med datamaskinen innebär inte att människan betraktas som en automat, inte heller att datamaskinen någonsin kan förväntas nå människans kommunikativa höjder. Genom att placera datorn på språkbrukarens plats ställer vi de lingvistiska frågorna på sin spets. Det är det som är den vetenskapliga poängen.

Det finns en viktig poäng till. Datorn får ju allt fler användningar i samhällslivet. Därigenom ökar behovet av kommunikation med den. Det är väsentligt att denna kommunikation sker på människans villkor.

Var ligger tyngdpunkten?

Låt oss först kasta en blick på modellen. Den ser ut ungefär så här. Språkliga yttringar kommer in till datamaskinen. Det gäller att avkoda dem – jag undviker ordet förstå. De skall i sin tur ge upphov till utdata, som kan vara av många slag, t ex en omskrivning, en konkret åtgärd, placering av språkgodset på rätt plats i en språkbank, ett svar på en fråga, en sammanfattning eller ett utkast till översättning.

Allt detta kräver naturligtvis kommunikativa resurser. Dem kan man ordna i fyra grupper. Med din tillåtelse betecknar jag dem med några facktermer och kommer strax in på vad de står för: lingvistisk kapacitet, encyklopedisk information, dynamiska faciliteter och algoritmer. Den lingvistiska kapaciteten gäller det lexikaliska systemet, det grammatiska systemet och språkbruket. Observera att data om språkbruket, användningen av språksystemet, inte är mindre viktiga än data om systemet självt. Man kan varken analysera eller generera naturligt språk på ett tillfredsställande sätt utan en uppsjö av information om hur och i vilken utsträckning språksystemets olika delar används.

Den encyklopediska information som krävs är systematiserad kunskap om världen. Vad kan exempelvis befinna sig ovanpå något annat? Fanns det verkligen bussar på Karl XII:s tid? Uppgiften är lindrigt talat stor och långsiktig. Den inbegriper flera olika vetenskapsgrenar. Detsamma gäller de dynamiska faciliteterna, som innefattar sådant som modeller för uppbyggande av kunskap genom perception och slutsatsdragning och strategier för konversation. Den forskning som bedrivs inom området artificiell intelligens är av stor betydelse härvidlag. Algoritmer slutligen behövs naturligtvis över hela fältet. Deras roll är ju att i detalj ange de procedurer som programmeringen bygger på.

Svaret på frågan om var tyngdpunkten ligger blir efter detta, att utforskningen av den lingvistiska kapaciteten och de därmed oskiljaktigt förenade algoritmerna är det centrala.

Människan har ju också en psykologisk dimension. Vad blir det av den?

Analogin med språkbrukaren framhäver i själva verket den aspekten. Att en modell har vad som brukar kallas psykologisk relevans får allt större betydelse. Den psykolingvistiska forskningen är också livaktig och följs med uppmärksamhet. Resultaten är emellertid ännu ganska motsägelsefulla och därför inte utan vidare tillämpliga. Men om det kan göras troligt att den ena av två i övrigt likvärdiga modeller är psykologiskt mera träffande än den andra, är den naturligtvis att föredra.

I vårt arbete på att komma åt de operativa enheterna i det lexikaliska betydel-

sesystemet har vi försökt att fånga de kategorier som kan antas vara relevanta för en observant och intresserad språkbrukare med tillgång till några vanliga nuspråkliga ordböcker.

Har detta synsätt stått klart från början?

Det har vuxit fram efterhand. Några saker har stått klara från början. Dit hör betydelsen av att undersöka autentiskt språk, att göra det på basis av teoretiskt grundade modeller och att utveckla datamaskinell metodik för det.

Vad kan datamaskinen redan nu göra av det vi har varit inne på?

Jag kanske får precisera din fråga på en viktig punkt. Datamaskinen gör inget annat än exakt det som vi programvägen har gett den möjlighet att göra. Det är alltså vi som verkar genom den. När detta är sagt, vill jag svara att vi under överskådlig tid får räkna med samarbete mellan maskinens programsystem och människan genom interaktiv databehandling. Successivt kan behovet av mänskligt ingripande under körningarna av allt att döma bli mindre. I flera avseenden kommer det förmodligen att bestå, såsom vid lösning av besvärliga flertydigheter, hantering av avsiktliga ordlekar av typen "Niagara är ett gränsfall", analys av texter där olika sociala och religiösa system interfererar med varandra, översättning av många slags texter osv. Å andra sidan finns det redan nu programsystem som kan utföra ganska goda morfologiska, syntaktiska och lexikaliska analyser, svara på frågor inom vissa områden osv.

Sedan hör det till saken att datorn också är ett redskap i själva forskningsprocessen. Den används nämligen för att bygga upp de kunskapsmängder och utveckla de system som krävs enligt språkbrukarmodellen. När det gäller den lingvistiska kapaciteten kan datamaskinen bland annat användas för experiment, t ex testning av språkregler, och insamling och bearbetning av lexikaliskt material. Den kan också användas som en stor och flexibel informationsbevarare.

Vilken roll spelar datateknikens utveckling?

Den är viktig på det sättet att den successivt gör tidigare omöjliga ting möjliga och tidigare tidsödande ting snabbt avklarade. Och utvecklingen har gått raskt. En stordator från mitten av 1950-talet som kostade 3 000 000 kronor kunde utföra 2000 instruktioner per sekund och hade en vikt av 7 ton. En mikrodator från mitten av 1970-talet som kostade 3000 kronor kunde utföra 200 000 instruktioner per sekund och hade en vikt av 7 kilogram.

För oss är minidatorn inte minst viktig. Den första modellen av en sådan lanserades 1960. Man har jämfört utvecklingen av minidatorn med utvecklingen av

en bil som Volkswagen för att konkretisera förhållandena. Om bilen hade haft motsvarande utveckling av prestanda och pris, så skulle den nu ha haft en topphastighet av 100 000 kilometer i timmen och kostat 10 kronor. Den avgörande slutsatsen av detta är, att vi inte skall låta oss hindra av för tillfället rådande teknologiska begränsningar vid utvecklingen av forskningsmetodiken på området. Datamaskinen är trots allt bara ett tredjedels sekel gammal.

Vilken utrustning har institutionen?

Vi har vad man kan kalla ett stort minidatorsystem som är speciellt utformat för behandling av språkligt material. Det kräver bland annat stor minneskapacitet, god strängbehandling, obegränsad typuppsättning och terminaler för interaktiv bearbetning. Forskarnas närkontakt med datamaskinen via textskärmsterminaler är av mycket stor vikt. Vi utnyttjar också flitigt ett par terminaler som är kopplade till Göteborgs Datacentral, där vi får tunga bearbetningar utförda.

Vilken inriktning har institutionens forskning för närvarande?

Frågan förutsätter helt rimligt att vi inte försöker täcka hela fältet. Man kan säga att datalingvistiska undersökningar av det lexikaliska systemet och av språkbruket utgör huvudområdena. Efter frekvensundersökningen som nu avslutas är det största projektet det som syftar till att lägga upp en omfattande svensk lexikalisk databas på modern lingvistisk grund och med det levande språket som källa. Genom denna förs utvecklingen av lexikon för datamaskinell analys och syntes vidare. Från databasen skall vi också generera en ordbok över modern svenska för allmänt bruk. Andra lexikaliska arbeten gäller ordlistor för invandrare och en studie av alla svenskars namn. Namn är ett rätt stort inslag i språkets lexikon. Under hösten 1979 utkommer Förnamnsboken.

Vi strävar efter att också arbeta med projekt som rör system och procedurer på den mera renodlat grammatiska sidan inom ramen för algoritmisk textanalys. En studie i program för morfologisk och syntaktisk analys är just avslutad.

Av flera skäl är språkmaterialet oftare skrivet än talat språk. En del av undersökningarna tar emellertid hänsyn till talspråk. Ett projekt arbetar helt med det. Undersökningen gäller vissa kommunikativa drag i talat vardagsspråk i Göteborg och deras relation till sociala faktorer.

Vad får forskningen för konsekvenser för teoribildningen?

Tillkomsten för första gången i historien av en symbolbehandlande maskin är naturligtvis en utmaning för lingvistiken. Detta understryks av att minneskapacitet, hastighet och kompakthet nu närmar sig den mänskliga hjärnans. I stället för

att begränsa oss till spekulationer om språkbeskrivningens utformning kan vi använda datamaskinen som redskap. Det visar sig att också starkt uppmärksammade teoretiska inriktningar råkar i svårigheter. Det som skymtar är en lingvistisk teori som en teori om analys och syntes av naturligt språk med hjälp av en uppsättning kommunikativa resurser inklusive algoritmer.

Det talas ibland om mjukdata. Vad blir det av dem?

I vissa sammanhang talar man mycket riktigt om mjukdata, som då gärna ställs i motsats till hårddata. Detta väcker lite av samma föreställningar som mjukvara och hårdvara inom databehandlingen, dvs program respektive maskinutrustning. Intressant är då att gränsen mellan dessa håller på att mjukas upp. Med hjälp av mikroteknik kan program nu monteras in som maskinkomponenter. Vi får hård mjukvara.

Som mjukdata betecknas exempelvis uppgifter om människors attityd till olika språkliga uttryck, t ex yrkesbeteckningar, människors upplevelse av sjukdom eller åldrande osv. Som hårddata betecknas då observerade språkliga frekvenser, resultat av laboratorieprov osv. I båda fallen gäller emellertid att uppgifterna måste svara mot vetenskapliga krav på noggrannhet i dokumentationen. De kan då principiellt också behandlas med likartade metoder. Orden mjukdata och hårddata representerar därför snarare olika vetenskapliga forskningsinriktningar än fundamentalt olika slag av data.

Kan man då också säga att epistemiska, fenomenologiska, holistiska och ontologiska frågeställningar inom språkvetenskapen är förenliga med forskningen på området?

Naturligtvis.

Är det riktigt att beteckna ämnet som tvärvetenskapligt?

Ja och nej. Beteckningen tvärvetenskap är inte alldeles klar. Vid dess sida står bland annat mångvetenskap som benämning på forskningsföretag som bygger på samverkan mellan olika vetenskaper. Sann tvärvetenskap innebär en sammansmältning av flera ämnen till en ny helhet. Detta är den tvärvetenskapliga paradoxen. När ett nytt ämne har bildats på det sättet, kan det ju inte längre med rätta kallas tvärvetenskapligt annat än från historisk synpunkt.

Vårt ämne har emellertid livliga förbindelser med andra ämnen. Det är också hjälpvetenskap för ämnen inom och utom språkvetenskapen och utnyttjar i sin tur andra ämnen som hjälpvetenskaper.

Vad gäller för utbildningen?

Som den är upplagd nu riktar den sig enbart till doktorander och utgår från grundexamen med tre terminers studier i allmän språkvetenskap eller motsvarande kunskaper. Studier i informationsbehandling är inte ett förkunskapskrav. Den som saknar sådana kunskaper får tillägna sig dem under utbildningen genom kursläsning och laborationer. Utöver doktorandutbildningen ger vi ibland orienteringskurser för olika grupper. Den verksamheten borde nog vidgas.

Vart vänder man sig om man vill utnyttja institutionens språkmaterial?

Då tar man kontakt med institutionens servicecorgan Logoteket. Det är en dynamisk språkbank som inrättades 1975 som en permanentning av en länge bedriven serviceverksamhet. Logoteket har vissa nationella uppgifter, nämligen att samla in och bevara datamaskinellt läsbara texter, att tillhandahålla sådana texter och vissa bearbetningar av dem, att bygga upp en ordbank och att ge råd på sitt område, allt i mån av resurser.

Textsamlingarna omfattar nu omkring 30 miljoner ord ur bland annat skönlitteratur, tidningar och författningar. Två speciella textmängder är särskilt aktuella. Den ena är riksdagens snabbprotokoll från arbetsåret 1978-79, tillsammans omkring fyra miljoner ord i en språkform någonstans mellan talspråk och skriftspråk. Den andra är Strindbergs samlade skrifter som skall kodas in och bearbetas här som ett led i kulturrådets nyutgivning. Omfånget uppskattas till mellan sex och sju miljoner ord.

Ordbanken innehåller ett par hundra tusen ord försedda med varierande uppgifter. Huvudinslag är frekvensordbokens olika bearbetningar och Svenska Akademiens ordlista. De inkommande texternas ordförråd tillförs successivt.

Konkordanser på mikrokort framställs systematiskt. Belägg på olika typer av ord och konstruktioner efterfrågas också. Avnämarna finns inom språkforskning och samhällsforskning, språkvård och journalistik, informationsbehandling och informationssökning, olika grenar av grafisk industri osv.

Hör arbetet kring språkbanken till den vetenskapliga verksamheten?

På flera sätt. Urvalet av språkmaterial och utformningen av databaserna, innehållsligt och tekniskt, är exempel på forskningsproblem. Men det finns skäl att anlägga ett vidare perspektiv. Inom en vetenskapsgren, och då inte minst språkforskningen, står vad man kunde kalla vetenskapsvården i centrum. Vad det gäller är teoribildning, metodutveckling och kunskapsgenerering. Ett annat betydelsefullt område är utbildningen, som ju är en förutsättning för att den vetenskapliga genomlysningen skall föras vidare. Ett tredje område har jag kallat resultatvård.

Här hör servicen hemma tillsammans med praktiska tillämpningar inom olika samhällssektorer. Forskarnas del i ansvaret för resultatens användning kommer också in i bilden.

En sak till. Intuitionen, vart tar den vägen?

I grunden finns det ingen motsättning, menar jag, mellan ett datalingvistiskt betraktelsesätt och en intuitiv uppläggning. Ytterst bör det nämligen vara så, att intuition är grundad på skarp observation. Förhåller det sig på det sättet, och min intuition får mig att tro det, blir ämnesområdets framtid inte mindre intressant.

Efter avslutat samtal förfogar man sig på klassiskt maner till caldarium, där armhävning och hett bad väntar.

STURE ALLÉN